Contrôle post hoc des faux positifs pour des hypothèses structurées

Guillermo Durand

LMO, Orsay

Joint work with Gilles Blanchard, Pierre Neuvial and Etienne Roquain
Séminaire de Probabilités et Statistique de l'IMAG

13/11/2023

Table of contents

Multiple testing setting

- Random data $X : (\Omega, \mathcal{T}, \mathbb{P}) \to (\mathcal{X}, \mathfrak{X})$ with unknown distribution $\mathcal{L}(X) \in \mathscr{P}$ a family of distributions
- m null hypotheses $H_{0,i} \subset \mathscr{P}$ on $\mathcal{L}(X)$
- $\mathcal{H}_0 = \{i : \mathcal{L}(X) \in H_{0,i}\}$: $i \in \mathcal{H}_0 \Leftrightarrow H_{0,i}$ is true
- m p-values $p_i = p_i(X)$ such that $p_i \succeq \mathcal{U}([0,1])$ if $i \in \mathcal{H}_0$
- Our object of interest: for every subset of hypotheses $S \subseteq \mathbb{N}_m$: $V(S) = |S \cap \mathcal{H}_0|$

Multiple testing setting

Toy setting, used for simulations

- Gaussian one-sided case: $X = (X_1, \ldots, X_m)$, $\mathcal{L}(X) \in \mathscr{P} = \{\mathcal{N}(\boldsymbol{\mu}, \mathrm{Id}_m) : \forall i \in \mathbb{N}_m, \mu_i \geq 0\}$
- We test, for all $i \in \mathbb{N}_m$, $H_{0,i} : \mu_i = 0$ versus $H_{1,i} : \mu_i > 0$.
- Formally, $H_{0,i} = \{\mathcal{N}(\boldsymbol{\mu}, \mathrm{Id}_m) \in \mathscr{P} : \mu_i = 0\}$
- $p_i(X) = p_i(X_i) = 1 - \Phi(X_i)$ with Φ the c.d.f. of $\mathcal{N}(0, 1)$

Multiple testing setting

Classical MT theory

- Form a rejection procedure $R : \mathcal{X} \to \mathcal{P}(\mathbb{N}_m)$ with a statistical guarantee on $V(R(X))$ no matter $\mathcal{L}(X)$
- $\text{FWER}(R) = \mathbb{P}(V(R(X)) > 0)$
  - Controlled by the famous Bonferroni procedure: $R_{Bonf}(X) = \left\{i : p_i(X) \leq \frac{\alpha}{m}\right\}$.
- FWER control too stringent for applications $\Rightarrow$ $\text{FDP}(R, X) = \frac{V(R(X))}{|R(X)| \vee 1}$ (difficult to control) or $\text{FDR}(R) = \mathbb{E}\left[\text{FDP}(R, X)\right]$.
  - FDP or FDR control $\Rightarrow$ allow for some false positives
  - Controlled under PRDS (or independence) by the Benjamini-Hochberg procedure [Benjamini and Yekutieli (2001)]
  - BH: let $\hat{k}_{BH} = \max\left\{k : p_{(k)}(X) \leq \frac{\alpha k}{m}\right\}$, then $R_{BH}(X) = \left\{i : p_i(X) \leq \frac{\alpha \hat{k}_{BH}}{m}\right\}$.

Exploratory analysis in multiple testing

Exploratory analysis: searching interesting hypotheses that will be cautiously investigated after.

Desired properties [Goeman and Solari (2011)]:

- Mildness: allows some false positives
- Flexibility: the procedure does not prescribe, but advise
- Post hoc: take decisions on the procedure after seing the data

[Goeman and Solari (2011)]

This **reverses the traditional roles** of the user and procedure in multiple testing. Rather than [...] to let the user choose the quality criterion, and to let the procedure return the collection of rejected hypotheses, the **user chooses the collection of rejected hypotheses freely**, and the multiple testing procedure returns the **associated quality criterion**.

Post hoc and replication crisis

Post hoc done wrong: *p*-hacking

- Pre-selecting variables that seem significant, exclude others
- Theoretical results no longer hold because the selection step is random
- Example: selecting the 1000 smallest *p*-values in a genetic study with 10^6 variants

- *p*-hacking may be one of the causes of the replication crisis
- Replication crisis: many published results non reproductible

$\Rightarrow$ need for exploratory analysis MT procedures with the above properties

Table of contents

Our goal: post hoc inference

Or simultaneous inference

Confidence bounds on any set of selected variables

A (post hoc) confidence bound is a random function

$$\widehat{V} : \mathcal{P}(\mathbb{N}_m) \to [\![0, m]\!]$$

such that:

$$\mathbb{P}\left(\forall S \subset \mathbb{N}_m, V(S) \leq \widehat{V}(S)\right) \geq 1 - \alpha. \quad (1)$$

$\widehat{V}$ depends on X and (1) has to be true no matter $\mathcal{L}(X)$.

- Hence for any selected $\widehat{S}$, $\mathbb{P}\left(V(\widehat{S}) \leq \widehat{V}(\widehat{S})\right) \geq 1 - \alpha$ holds
- Also an FDP bound: $\mathbb{P}\left(\forall S \subset \mathbb{N}_m, \mathrm{FDP}(S) \leq \widehat{V}(S)/|S|\right) \geq 1 - \alpha$
- $\Longrightarrow$ allows construction of sets with bounded FDP
- Originates from [Genovese and Wasserman (2006) and Meinshausen (2006)]
- A guarantee over any selected set instead of a rejected set, advise some $\widehat{S}$ instead of prescribe one R : the MT paradigm is reversed

BNR technology

[Blanchard et al. (2020)]

Key concept: reference family

- $\mathfrak{R} = (R_k, \zeta_k)_{k \in K}$ (random) such that Joint Error Rate (JER) control:

$$\text{JER}(\mathfrak{R}) = \mathbb{P}\left(\exists k, |R_k \cap \mathcal{H}_0| > \zeta_k\right) \leq \alpha. \tag{2}$$

  $\mathfrak{R}$ depends on X and (2) has to be true no matter $\mathcal{L}(X)$.
- Conversely, $\mathbb{P}\left(\forall k, V(R_k) \leq \zeta_k\right) \geq 1 - \alpha$
- Confidence bound only on the members of $\mathfrak{R}$
- $\Longrightarrow$ Derivation of a global confidence bound by interpolation

BNR technology

[Blanchard et al. (2020)]

Idea: we get the following info on $\mathcal{H}_0$:

$$\mathcal{H}_0 \in \mathcal{A}(\mathfrak{R}) = \{A \in \mathcal{P}(\mathbb{N}_m) : \forall k, |R_k \cap A| \leq \zeta_k\}.$$

Two different bounds

- $V^*_{\mathfrak{R}}(S) = \max\{|S \cap A| : A \in \mathcal{A}(\mathfrak{R})\}$ optimal but hard to compute (possibly NP)
- $\overline{V}_{\mathfrak{R}}(S) = \min_k (\zeta_k + |S \setminus R_k|) \wedge |S|$ easier to compute, $\geq V^*_{\mathfrak{R}}(S)$

BNR technology

Family construction

- In [Blanchard et al. (2020)], $\zeta_k = k - 1$ always, and $R_k = \{i : p_i < t_k\}$ such that JER control.
- Example: $t_k = \alpha k/m$ (Simes inequality) if p-values PRDS.
- $\Rightarrow$ JER control becomes "simultaneous k-FWER control"

Table of contents

Spatial structure

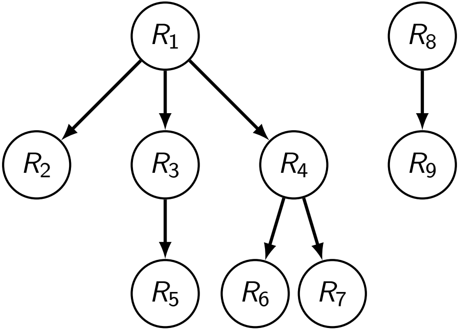
Informal assumption

The signal is localized in some spatially structured regions, with, possibly, different levels, that we can access to with previous information (e.g. active SNPs into genes into chromosomes)

- Accordingly, find adapted new reference families, using those regions
- We want $V^*_{\mathfrak{R}}$ to be easy to compute
- Our approach: deterministic R_k's capturing spatial hierarchy, estimate the true nulls inside them (i.e. ζ_k random)
  - the opposite of [Blanchard et al. (2020)]

Forest structure

- $\forall k, k' \in \mathcal{K}, \; R_k \cap R_{k'} \in \{R_k, R_{k'}, \varnothing\}$
- Connected components are trees:



- Accommodates to different levels of signal localization through the different depths of the nodes
- Includes nested families or totally disjoint families

New interpolation bounds

Goal: compute $V^*_{\mathfrak{R}}$ easily with forest structure

- Recall $\overline{V}_{\mathfrak{R}}(S) = \min_{k\in\mathcal{K}} \left(\zeta_k \wedge |S \cap R_k| + |S \setminus R_k|\right)$

Definition

For any $q \leq K = |\mathcal{K}|$,

$$\widetilde{V}^q_{\mathfrak{R}}(S) = \min_{Q\subset\mathcal{K}, |Q|\leq q} \left(\sum_{k\in Q} \zeta_k \wedge |S \cap R_k| + \left| S \setminus \bigcup_{k\in Q} R_k \right| \right).$$

Property

$$V^*_{\mathfrak{R}}(S) \leq \widetilde{V}^K_{\mathfrak{R}}(S) \leq \widetilde{V}^{K-1}_{\mathfrak{R}}(S) \leq \cdots \leq \widetilde{V}^2_{\mathfrak{R}}(S) \leq \widetilde{V}^1_{\mathfrak{R}}(S) = \overline{V}_{\mathfrak{R}}(S)$$

Main results

Compute $V^*_{\mathfrak{R}}$ easily with forest structure

Theorem

$$V^*_{\mathfrak{R}}(S) = \widetilde{V}^K_{\mathfrak{R}}(S)$$

Even better,

$$V^*_{\mathfrak{R}}(S) = \widetilde{V}^{\ell}_{\mathfrak{R}}(S),$$

with ℓ = number of leaves = max number of disjoint sets

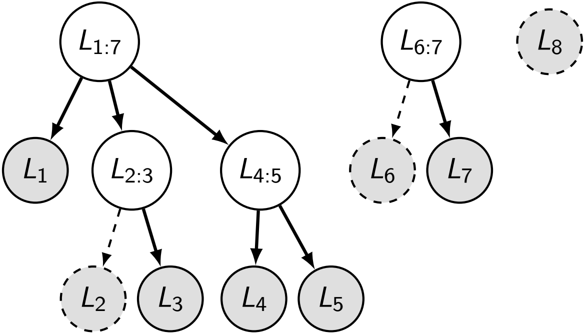
Corollary

$\ell = 1$ for nested families and a property in BNR is recovered

Forest structure

Property (completion)

- Each forest structure can be completed to includes all leaves
- Regions are disjoint unions of leaves: $R_k = \bigcup_{\ell=i}^{j} L_\ell = L_{i:j}$
- For an added leaf L_ℓ, just state $\zeta_\ell = |L_\ell|$



Main results

Compute $V^*_{\mathfrak{R}}$ easily with forest structure

Corollary (derived from the proof by construction)

There is a simple and efficient algorithm to compute $\widetilde{V}^K_{\mathfrak{R}}(S)$ if $\mathfrak{R}$ is complete ($O(Hm)$ complexity).

Lemma

Completing the family does not change $V^*_{\mathfrak{R}}$ and $\widetilde{V}^K_{\mathfrak{R}}$.

Corollary

There is a simple algorithm to compute $V^*_{\mathfrak{R}}(S)$ in any case by completing the family first.

Note: all of the above does not depend on the choice of the ζ_k and works for random R_k.

Forest algorithm

Computation of $V^*_{\mathfrak{R}}(S)$

Data: $\mathfrak{R} = (L_{i:j}, \zeta_{i,j})_{(i,j)\in\mathcal{K}}$ and $S \subset \mathbb{N}_m$

$\mathfrak{R} \longleftarrow \mathfrak{R}^{\oplus}$; $\mathcal{K} \longleftarrow \mathcal{K}^{\oplus}$; $n \longleftarrow$ final number of leaves (completion)

$Vec \longleftarrow (0, \ldots, 0) \in \mathbb{R}^n$ (initialisation)

$H \longleftarrow$ maximum depth

for (i, j) *at depth* H **do**

    $Vec_i \longleftarrow \zeta_{i:j} \wedge |S \cap R_{i:j}|$

end

for $h \in \{H-1, \ldots, 1\}$ **do**

    for (i, j) *at depth* h **do**

        $Succ_{i:j} \longleftarrow \{(i', j')$ at depth $h+1 : R_{i':j'} \subset R_{i:j}\}$

        $Vec_i \longleftarrow \min\left(\zeta_{i:j} \wedge |S \cap R_{i:j}|, \sum_{(i',j')\in Succ_{i:j}} Vec_{i'}\right)$

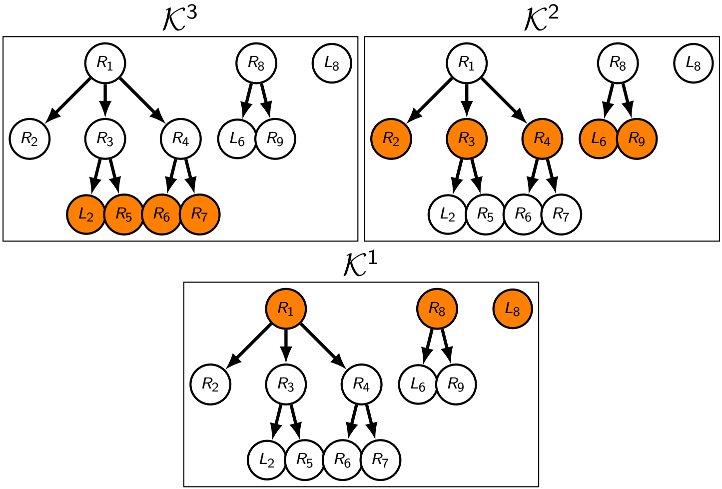
        $Vec_\ell \longleftarrow 0$ for all $i+1 \leq \ell \leq j$

    end

end

return $\sum_{i=1}^{n} Vec_i$.

Forest algorithm



Problem

- We often want to compute $V^*_{\mathfrak{R}}(S)$ for a path S_t, $1 \leq t \leq m$
- For example $S_t = \{\text{the indexes of the } t \text{ smallest } p\text{-values}\}$
- The above algorithm becomes slow
- Is there a may to leverage the fact that we add one p-value at a time to update $V^*_{\mathfrak{R}}(S_t)$ quickly?
- YES!
- And we can also get the partition that realizes the min

$$V^*_{\mathfrak{R}}(S_t) = \min_{Q \subset \mathcal{K}^{\oplus}, Q \text{ partition}} \left(\sum_{k \in Q} \zeta_k \wedge |S_t \cap R_k| \right)$$

NEW algorithm

Fast computation of a path $(V^*_{\mathfrak{R}}(S_t))_{1 \leq t \leq m}$

```
V_0 ⟵ 0, K^- ⟵ {k ∈ K : ζ_k = 0}, ∀k ∈ K, η_k ⟵ 0
for t = 1, ..., m do
    if i_t ∈ ⋃_{k∈K^-} R_k then
        V_t ⟵ V_{t-1}
    end
    else
        for h = 1, ..., h_max(t) do
            find k^(t,h) at depth h such that i_t ∈ R_{k^(t,h)}
            η_{k^(t,h)} ⟵ η_{k^(t,h)} + 1
            if η_{k^(t,h)} < ζ_k then
                pass
            end
            else
                K^- ⟵ K^- ∪ {k^(t,h)}
                break the loop
            end
        end
        V_t ⟵ V_{t-1} + 1
    end
end
return (V_t)_{1≤t≤m}
```

True nulls estimation inside regions

That is, ζ_k computation

- K deterministic regions, let $C = \sqrt{\frac{1}{2}\log\left(\frac{K}{\alpha}\right)}$
- $\zeta_k = |R_k| \wedge \min_{t\in[0,1)} \left\lfloor \frac{C}{2(1-t)} + \left(\frac{C^2}{4(1-t)^2} + \frac{\sum_{i\in R_k} \mathbb{1}\{p_i > t\}}{1-t} \right)^{1/2} \right\rfloor^2$
- Comes from handling the DKWM inequality [Dvoretzky et al. (1956) and Massart (1990)]
  - Requires independence!
- Replace $\min_{t\in[0,1)}$ and t above by $\min_{0\leq\ell\leq s}$ and $p_{(\ell)}$ for practical usage $\Longrightarrow$ computation of all ζ_k is also $O(Hm)$ complex
- α/K instead of α in C: union bound for JER control
- Dependence on α (and to K!) only through a log
- $\zeta_k > 0$ (entry cost)

DKWM use

DKWM :

$$v^{-1}\sum_{i=1}^{v}\mathbb{1}\{U_i > t\} - (1-t) \geq -\sqrt{\log(1/\lambda)/(2v)}, \ \forall t \in [0,1],$$

with probability at least $1-\lambda$, for $U_1,\dots,U_v$ i.i.d. $\mathcal{U}([0,1])$.

2nd degree polynom manipulation yields:

$$v \leq \min_{t\in[0,1)} \left(\frac{\sqrt{\log(1/\lambda)/2}}{2(1-t)} + \left(\frac{\log(1/\lambda)/2}{4(1-t)^2} + \frac{\sum_{i=1}^{v}\mathbb{1}\{U_i > t\}}{1-t} \right)^{1/2} \right)^2$$

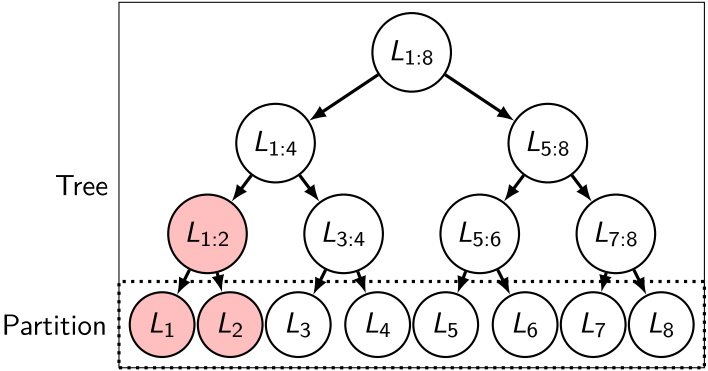
- $v = |R_k \cap \mathcal{H}_0|$, $\lambda = \alpha/K$
- $\sum_{i\in R_k}\mathbb{1}\{p_i(X) > t\}$ dominates $\sum_{i=1}^{v}\mathbb{1}\{U_i > t\}$ by independence

Table of contents

Comparison of 3 bounds

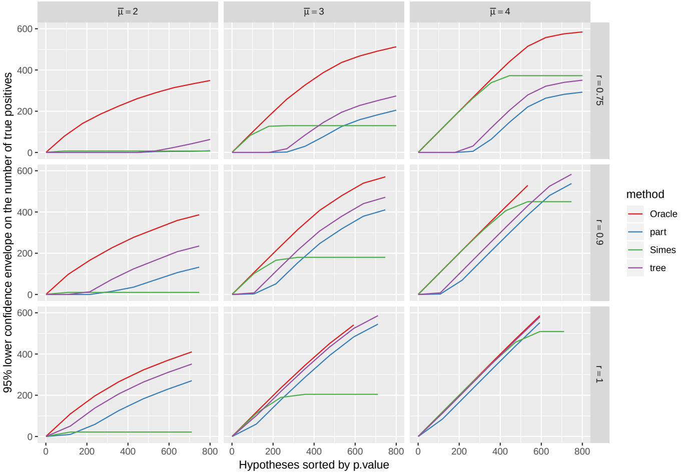
Simes bound of BNR, and 2 new

- V_{tree} and V_{part}: complete binary tree or only the leaves partition
- Signal in adjacent leaves, expectation of V_{tree} good despite worst K
- Parameters: signal $\bar{\mu}$ and signal proportion in active leaves r



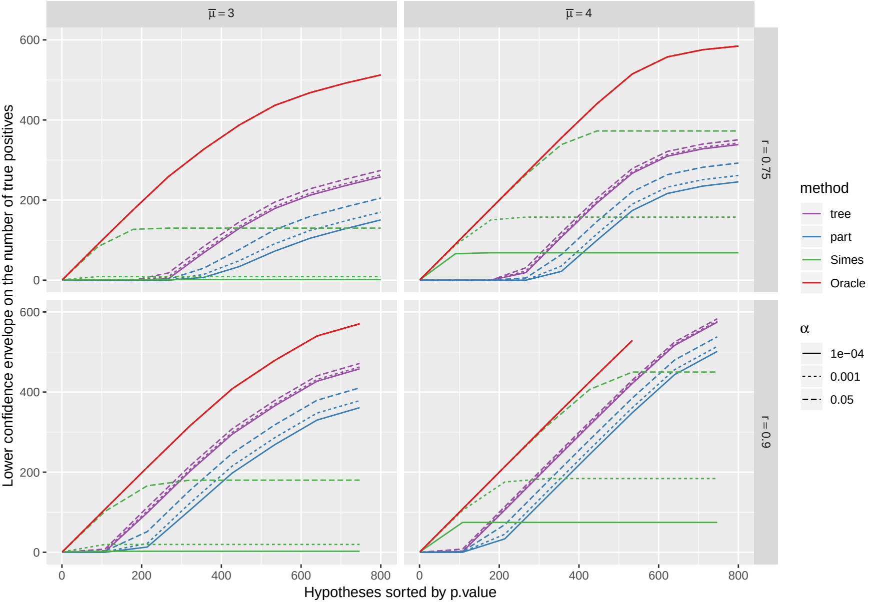
Comparison of 3 bounds

- S_t = the t-th smallest p-values
- Simes better for large signal, but new bounds better for large r, *generally*
- "Plateau" effect for Simes for large x
- V_{tree} better than V_{part} as expected, despite worst union bound constant



Comparison of 3 bounds

Influence of α



New hybrid bound suggested by the simulations

- $V_{\mathsf{hybrid}}^{\gamma}(\alpha, S) = \min\left(V_{\mathsf{Simes}}((1-\gamma)\alpha, S), V_{\mathrm{tree}}(\gamma\alpha, S)\right)$
- $\gamma = 0.02$: favors Simes, not a problem because V_{tree} is little sensitive to small α

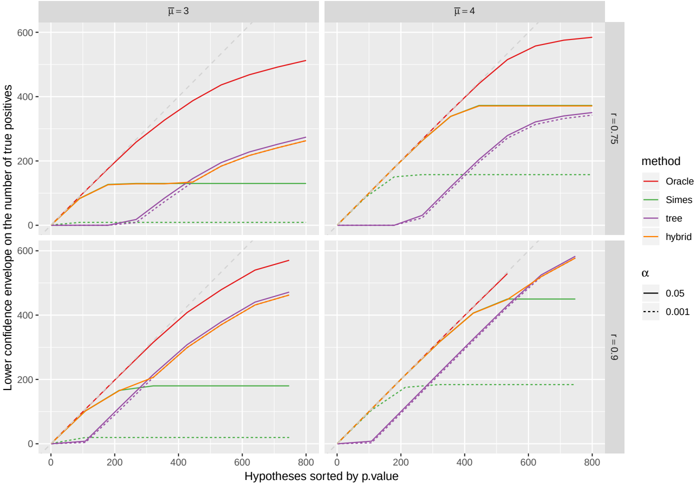
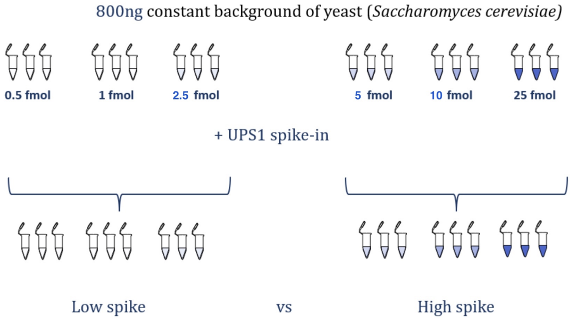


Table of contents

Application

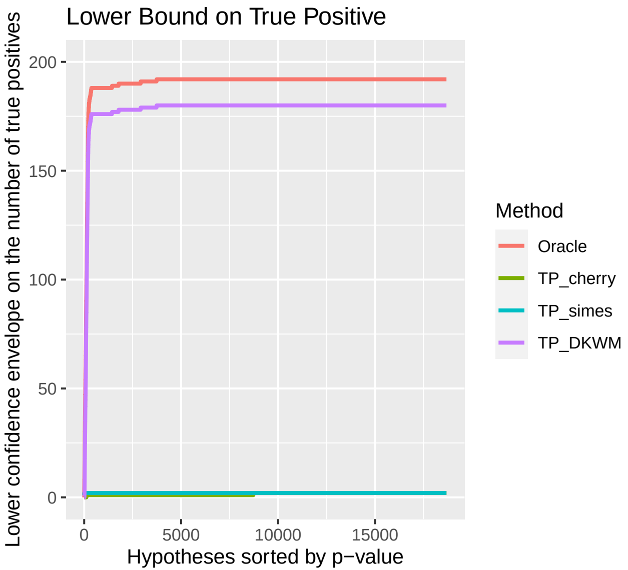
Proteomics data

- Joint work with Marie Chion, Alexandre Perrin, Auriane Gabaut, Mélina Gallopin, Romain Périer
- Data from [Chion et al. (2022)]



- Controlled(-ish) experiment : $H_{0,i}$ is known for all i!
- One-sided p-values from a mean comparison test (Welch test)

Application



Application

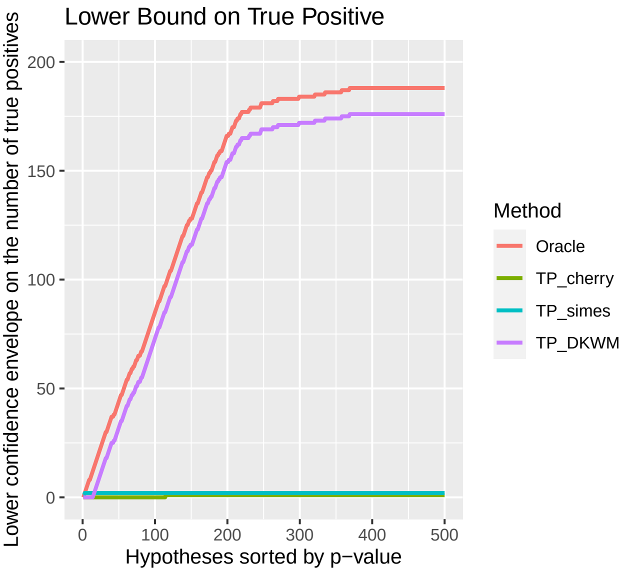


Table of contents

Conclusion

New confidence bounds that exploit the signal localization to improve on existing bounds, with an acceptable computation time

Limitations:

- DKWM inequality involves independence
- The chosen ζ_k can't reject a whole subset (including individual hypotheses)
- The R_k have to be fixed before seeing the data (not post hoc!)
- The union bound correction chosen may induce conservativeness

Published paper in Scandinavian Journal of Statistics (2020) [Durand et al. (2020)]
Also on arXiv: 1807.01470
R package available on github: sansSouci

Next steps

- Let $\phi(\cdot,\cdot,\cdot)$ a FWER-controlling test,
  - $\zeta_k = L_k\left(\frac{\alpha}{K}\right) = |R_k| - \left|\phi\left(X, \frac{\alpha}{K}, R_k\right)\right|$ is valid
  - No independence required if ϕ doesn't require it
  - $\zeta_k = 0$ doable $\iff V^*(S) = 0$ doable
- $\implies$ alternative L_k to DKWM. Also : permutation-based L_k? Concentration inequalities for (weakly) dependent variables? Estimators of [Chen (2019)]?
- Adaptation to the heterogeneous discrete setting?
- Bayesian post hoc with ℓ-values? With HMM? [Perrot-Dockès et al. (2023)]
- Learn the regions with a training set? [Blain et al. (2022)] already learn BNR t_k.
- Applications to neuroimagery [Vesely et al. (2021)]

Bibliography I

Benjamini, Yoav and Daniel Yekutieli (2001). “The control of the false discovery rate in multiple testing under dependency”. In: *Annals of statistics*, pp. 1165–1188.

Blain, Alexandre, Bertrand Thirion, and Pierre Neuvial (2022). “Notip: Non-parametric true discovery proportion control for brain imaging”. In: *NeuroImage* 260, p. 119492. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2022.119492. URL: https://www.sciencedirect.com/science/article/pii/S1053811922006085.

Blanchard, Gilles, Pierre Neuvial, and Etienne Roquain (2020). “Post hoc confidence bounds on false positives using reference families”. In: *The Annals of Statistics* 48.3, pp. 1281 –1303. DOI: 10.1214/19-AOS1847. URL: https://doi.org/10.1214/19-AOS1847.

Chen, Xiongzhi (2019). “Uniformly consistently estimating the proportion of false null hypotheses for composite null hypotheses via Lebesgue-Stieltjes integral equations”. In: *arXiv preprint arXiv:1906.10246*.

Chion, Marie, Christine Carapito, and Frédéric Bertrand (2022). “Accounting for multiple imputation-induced variability for differential analysis in mass spectrometry-based label-free quantitative proteomics”. In: *PLoS Computational Biology* 18.8, e1010420.

Durand, Guillermo et al. (2020). “Post hoc false positive control for structured hypotheses”. In: *Scandinavian journal of Statistics* 47.4, pp. 1114–1148.

Dvoretzky, Aryeh, Jack Kiefer, and Jacob Wolfowitz (1956). “Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator”. In: *The Annals of Mathematical Statistics*, pp. 642–669.

Bibliography II

Genovese, Christopher R and Larry Wasserman (2006). "Exceedance control of the false discovery proportion". In: *Journal of the American Statistical Association* 101.476, pp. 1408–1417.

Goeman, Jelle J and Aldo Solari (2011). "Multiple testing for exploratory research". In: *Statistical Science*, pp. 584–597.

Marcus, Ruth, Peritz Eric, and K Ruben Gabriel (1976). "On closed testing procedures with special reference to ordered analysis of variance". In: *Biometrika* 63.3, pp. 655–660.

Massart, Pascal (1990). "The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality". In: *The Annals of Probability*, pp. 1269–1283.

Meinshausen, Nicolai (2006). "False discovery control for multiple tests of association under general dependence". In: *Scandinavian Journal of Statistics* 33.2, pp. 227–237.

Perrot-Dockès, Marie et al. (2023). "Selective inference for false discovery proportion in a hidden Markov model". In: *TEST*, pp. 1–27.

Vesely, Anna, Livio Finos, and Jelle J Goeman (2021). "Permutation-based true discovery guarantee by sum tests". In: *arXiv preprint arXiv:2102.11759*.

Closed testing for post hoc inference

Designed for FWER control [Marcus et al. (1976)]

- Form $H_{0,I} = \bigcap_{i \in I} H_{0,i}$ $\forall I \subset \mathbb{N}_m$: all intersection hypotheses
- Have a collection of α level local tests ϕ_I
- Examples:
  - Bonferroni test $\phi_I = 1$ if $\exists i \in I : p_i \leq \alpha/|I|$
  - Simes test $\phi_I = 1$ if $\exists i \in I : p_{(i:I)} \leq \alpha i/|I|$ (under PRDS)
- Test $H_{0,I}$ only if all $H_{0,J}$, $J \supseteq I$, are rejected
- Reject the individual hypotheses $H_{0,i}$ such that $H_{0,\{i\}}$ has been rejected that way
- Then FWER(Closed testing) $\leq \alpha$

Closed testing for post hoc inference

[Goeman and Solari (2011)]

Main idea

The closed testing provides more information than just the individual rejects:

- Let $\mathcal{X}$ the set of all I such that we rejected $H_{0,I}$
- Simultaneous guarantee over all $H_{0,I}$, $I \in \mathcal{X}$:

$$\mathbb{P}\left(\forall I \in \mathcal{X}, H_{0,I} \text{ is false}\right) \geq 1 - \alpha$$

Confidence bound derivation:

- $V_{\mathrm{GS}}(S) = \max_{\substack{I \subseteq S \\ I \notin \mathcal{X}}} |I|$ is a confidence bound because

$$\begin{aligned}\exists S, |S \cap \mathcal{H}_0| > V_{\mathrm{GS}}(S) &\Longrightarrow \exists S, S \cap \mathcal{H}_0 \in \mathcal{X} \\ &\quad \text{but } H_{0,S\cap\mathcal{H}_0} \text{ is true} \\ &\Longrightarrow \exists I \in \mathcal{X}, H_{0,I} \text{ is true}\end{aligned}$$

- $V_{\mathrm{GS}}(S) = V^*_{\mathfrak{R}}(S)$ with $\mathfrak{R} = (I, |I| - 1)_{I \in \mathcal{X}}$